

**HHS PUBLIC ACCESS**

Author manuscript

*Surv Methodol.* Author manuscript; available in PMC 2018 March 06.

Published in final edited form as:

*Surv Methodol.* 2015 December ; 41(2): 389–401.**Optimum allocation for a dual-frame telephone survey****Kirk M. Wolter,**

NORC at the University of Chicago, 55 East Monroe Street, Suite 3000, Chicago, IL 60603

**Xian Tao,**

NORC at the University of Chicago, 55 East Monroe Street, Suite 3000, Chicago, IL 60603

**Robert Montgomery, and**

NORC at the University of Chicago, 55 East Monroe Street, Suite 3000, Chicago, IL 60603

**Philip J. Smith**

Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Disease, Immunization Services Division, MS A-19, 1600 Clifton Road, NE, Atlanta, GA 30333

**Abstract**

Careful design of a dual-frame random digit dial (RDD) telephone survey requires selecting from among many options that have varying impacts on cost, precision, and coverage in order to obtain the best possible implementation of the study goals. One such consideration is whether to screen cell-phone households in order to interview cell-phone only (CPO) households and exclude dual-user household, or to take all interviews obtained via the cell-phone sample. We present a framework in which to consider the tradeoffs between these two options and a method to select the optimal design. We derive and discuss the optimum allocation of sample size between the two sampling frames and explore the choice of optimum  $p$ , the mixing parameter for the dual-user domain. We illustrate our methods using the *National Immunization Survey*, sponsored by the Centers for Disease Control and Prevention.

**Keywords**Dual-frame surveys; Optimum allocation; Sample design; *National Immunization Survey***1 Introduction**

Modern random digit dial (RDD) telephone surveys in the U.S. use two samples: a sample of landlines and a sample of cell-phone lines. Wolter, Smith and Blumberg (2010) provide the statistical foundations for such dual-frame telephone surveys. The present article builds on that work and demonstrates the considerations and statistical methods for allocating the total survey resources to the two sampling frames.

Because it is less costly on a per-unit basis and has a longer history of use, the landline sample is often the larger sample and the survey interview is attempted for all respondents in

this sample. The interviewing protocol for the smaller cell-phone sample is configured in one of two ways: (1) attempt to complete the survey interview for all responding persons, or (2) conduct a brief screening interview to ascertain the telephone status of the respondent, and then attempt to complete the survey interview only for respondents whose telephone status is classified as cell-phone-only (CPO) (i.e., respondents who report in the screening interview that they do not have a working landline in their household). (Within the screening approach there are variations, such as interviewing both CPO respondents and others who report that there is a landline in the household but they are not reachable through the landline.) As the size of the landline-only (LLO) population (i.e., persons who have a working landline telephone in the household but do not have access to a cell phone) declines over time (Blumberg and Luke 2010), survey statisticians may consider new designs in which the cell-phone sample is the larger sample and all respondents are interviewed, while the interviewing protocol for the smaller landline sample calls for screening or taking all respondents. Yet in this article, we focus on the prevailing circumstances in the last several years in which the cell-phone sample is typically the smaller sample and a take-all or screening protocol is used for respondents in this sample.

We shall develop the methods for optimum allocation under ideal assumptions that the sample sizes refer to completed cases (i.e., no nonresponse); that there is essentially a one-to-one relationship between the sampling units (telephone numbers) and the analytical units (e.g., households) in the landline population; that there is essentially a one-to-one relationship between the sampling units and the analytical units in the cell-phone population; and that all units in the target population are included in at least one of the two sampling frames. Given these assumptions, each and every specific analytic unit is linked to a landline, a cell-phone line, or both a landline and a cell-phone line, and is linked to at most one landline and at most one cell-phone line.

Most of the previous literature on dual-frame surveys studies estimation procedures rather than the question of allocation of the sample size to the various sampling frames, including Hartley (1962, 1974); Fuller and Burmeister (1972); Skinner and Rao (1996); and Lohr and Rao (2000, 2006). Biemer (1984) and Lepkowski and Groves (1986) looked at allocation when one frame is a subset of the other frame, as might be the case with an area sample supplemented by a special list.

To begin, we establish our notation and assumptions. Let  $U^A$  be the landline population and  $U^B$  the cell-phone population. The overall population of interest is  $U = U^A \cup U^B$ . Some units have both a landline and a cell phone (the dual-user population), while others have only a landline (the LLO population) or only a cell phone (the CPO population), and thus the two populations overlap as follows:  $U^{ab} = U^A \cap U^B$ ,  $U^a = U^A - U^{ab}$ , and  $U^b = U^B - U^{ab}$ .  $U^a$  is the LLO domain,  $U^b$  is the CPO domain, and  $U^{ab}$  is the dual-user domain. The population sizes are  $N_A = \text{card}(U^A)$ ,  $N_B = \text{card}(U^B)$ ,  $N_{ab} = \text{card}(U^{ab})$ ,  $N_a = \text{card}(U^a)$ , and  $N_b = \text{card}(U^b)$ . We denote the proportions in the overlap (or dual-user) population by  $\alpha = N_{ab}/N_A$  and  $\beta = N_{ab}/N_B$ .

Let  $s_A$  be a simple random sample without replacement selected from  $U^A$ , let  $s_B$  be a simple random sample without replacement selected from  $U^B$ , and let  $n_A = \text{card}(s_A)$  and  $n_B =$

$\text{card}(s_B)$  be the sample sizes (i.e., completed interviews). We assume that domain membership ( $a$ ,  $ab$ ,  $b$ ) is not known at the time of sampling.

Let  $Y_i$  be a variable of interest for the  $i^{\text{th}}$  unit in the overall population. The population domain means and variance components are denoted by  $\bar{Y}_A$ ,  $\bar{Y}_B$ ,  $\bar{Y}_{ab}$ ,  $\bar{Y}_a$ ,  $\bar{Y}_b$ ,  $S_A^2$ ,  $S_B^2$ ,  $S_{ab}^2$ ,  $S_a^2$ , and  $S_b^2$ . We take the goal of the survey to be the estimation of the overall population total  $Y$ .

In what follows, we derive the optimum allocation given the take-all protocol and the screening protocols in Section 2 and Section 3, respectively. Section 4 compares the two protocols in terms of efficiency and cost and attempts to provide guidance about the circumstances under which each protocol is better. The section also explores the optimum choice of a mixing parameter  $p$ , which is used to combine the estimators from the two samples ( $s_A \cap U^{ab}$  and  $s_B \cap U^{ab}$ ) that represent the dual-user population. Section 5 applies the methods to the *National Immunization Survey*, a large dual-frame telephone survey sponsored by the Centers for Disease Control and Prevention (CDC). The article closes with a brief summary in Section 6.

## 2 Take-all protocol

In the take-all protocol, one conducts survey interviews for all units in both samples  $s_A$  and  $s_B$ . Therefore, variable data collection costs can be approximated by the model

$$C_{TA} = c_A n_A + c_B n_B, \quad (2.1)$$

where  $c_A$  is the cost per completed interview in sample  $s_A$  and  $c_B$  is the cost per completed interview in sample  $s_B$ . The expected numbers of survey interviews in the cell-phone sample are  $(1 - \beta)n_B$  CPO units and  $\beta n_B$  dual-user units.

The unbiased estimator of the population total (Hartley 1962) is given by

$$\ddot{Y} = \hat{Y}_a + p\hat{Y}_{ab} + q\hat{Y}_{ba} + \hat{Y}_b, \quad (2.2)$$

where  $p$  is a mixing parameter,  $q = 1 - p$ ,  $\hat{Y}_a = (N_A/n_A)y_a$  is an estimator of the LLO total,  $\hat{Y}_{ab} = (N_A/n_A)y_{ab}$  is an estimator of the dual-user total derived from the landline sample,  $\hat{Y}_{ba} = (N_B/n_B)y_{ba}$  is an estimator of the dual-user total derived from the cell-phone sample,  $\hat{Y}_b = (N_B/n_B)y_b$  is an estimator of the CPO total,  $y_a$  is the sum of the variable of interest for the observations in  $s_A$  and in domain  $U^a$ ,  $y_{ab}$  is the sum of the variable of interest for the observations in  $s_A$  and in domain  $U^{ab}$ ,  $y_{ba}$  is the sum of the variable of interest for the observations in  $s_B$  and in domain  $U^{ab}$ , and  $y_b$  is the sum of the variable of interest for the observations in  $s_B$  and in domain  $U^b$ . We examine the choice of  $p$  in Section 4.

Given fixed  $p$ , we find that the variance of  $\ddot{Y}$  is

$$\text{Var} \{ \ddot{Y} \} = N^2 \left( \frac{Q_A^2}{n_A} + \frac{Q_B^2}{n_B} \right), \quad (2.3)$$

where  $W_A = N_A/N$ ,  $W_B = N_B/N$ ,

$$Q_A^2 = W_A^2 \left\{ (1-\alpha) S_a^2 + \alpha p^2 S_{ab}^2 + \alpha (1-\alpha) (\bar{Y}_a - p \bar{Y}_{ab})^2 \right\},$$

and

$$Q_B^2 = W_B^2 \left\{ (1-\beta) S_b^2 + \beta q^2 S_{ab}^2 + \beta (1-\beta) (\bar{Y}_b - q \bar{Y}_{ab})^2 \right\}.$$

The classical optimum allocation of the total sample to the two sampling frames (Cochran 1977) is defined by

$$\begin{aligned} n_{A,opt} &= \frac{K Q_A}{\sqrt{c_A}} \\ n_{B,opt} &= \frac{K Q_B}{\sqrt{c_B}}, \end{aligned} \quad (2.4)$$

where  $K$  is a constant that depends upon whether the objective of the allocation is to minimize cost subject to a constraint on variance, or to minimize variance subject to a constraint on cost. The minimum variance subject to fixed cost  $C_{TA}$  is given by

$$\min [\text{Var} \{ \ddot{Y} \}] = \frac{(\sqrt{c_A} Q_A + \sqrt{c_B} Q_B)^2}{C_{TA}}, \quad (2.5)$$

while the minimum cost subject to fixed variance  $V_0$  is

$$\min [C_{TA}] = \frac{(\sqrt{c_A} Q_A + \sqrt{c_B} Q_B)^2}{V_0}. \quad (2.6)$$

### 3 Screening protocol

In the screening protocol, one conducts survey interviews for all units in the landline sample  $s_A$ . One conducts screening interviews (for telephone status) for all units in the cell-phone sample  $s_B$  and then conducts the survey interviews only for the units that screen-in as CPO. Therefore, expected data collection costs arise according to the model

$$\begin{aligned} C_{SC} &= c_A n_A + c'_B \beta n_B + c''_B (1-\beta) n_B \\ &= c_A n_A + c'''_B n_B, \end{aligned} \quad (3.1)$$

where  $c'_B$  is the cost per completed screener (to ascertain telephone status) in sample  $s_B$ ,  $c''_B$  is the cost per completed screener and interview in sample  $s_B$ , and  $c'''_B = c'_B \beta + c''_B (1-\beta)$ . In this notation,  $n_A$  is the number of survey interviews completed amongst landline respondents and  $n_B$  is the number of completed interviews (telephone screener only for non-CPO respondents, and screener plus survey interview for CPO respondents) amongst cell-phone respondents. That is, the expected total number of completed survey interviews is  $n_A + (1 - \beta)n_B$ .

The unbiased estimator of the overall population total is

$$\hat{Y} = \hat{Y}_A + \hat{Y}_b, \quad (3.2)$$

where  $\hat{Y}_A = (N_A/n_A)y_A$ ,  $\hat{Y}_b = (N_B/n_B)y_b$ , and  $y_A = y_a + y_{ab}$ . The variance of the estimator is

$$\text{Var} \{ \hat{Y} \} = N^2 \left( \frac{R_A^2}{n_A} + \frac{R_B^2}{n_B} \right), \quad (3.3)$$

where

$$R_A^2 = W_A^2 S_A^2$$

and

$$R_B^2 = W_B^2 S_b^2 \left\{ 1 - \beta + \beta (1 - \beta) \frac{\bar{Y}_b^2}{S_b^2} \right\}.$$

The optimal allocation of the total sample is

$$n_{A, opt} = LR_A / \sqrt{c_A}$$

$$n_{B, opt} = LR_B / \sqrt{c_B'''},$$

where  $L$  is a constant that depends on the fixed constraint: cost or variance. The minimum variance subject to fixed cost is given by

$$\min [\text{Var} \{ \hat{Y} \}] = \frac{(\sqrt{c_A} R_A + \sqrt{c_B'''} R_B)^2}{C_{SC}}, \quad (3.4)$$

and the minimum cost subject to fixed variance is

$$\min [C_{SC}] = \frac{(\sqrt{c_A} R_A + \sqrt{c_B'''} R_B)^2}{V_0}. \quad (3.5)$$

## 4 Comparing the take-all and screening protocols

We compare the take-all and screening protocols to establish which is the less costly or more efficient. Such a comparison can provide practical guidance to planners of future dual-frame telephone surveys.

### 4.1 Comparing the minimum variances and costs

Given either fixed cost or fixed variance, efficiency can be assessed in terms of the ratio

$$E = \frac{\min [\text{Var} \{ \hat{Y} \}]}{\min [\text{Var} \{ \hat{Y} \}]} = \frac{\min [C_{SC}]}{\min [C_{TA}]} = \frac{(\sqrt{c_A} R_A + \sqrt{c_B'''} R_B)^2}{(\sqrt{c_A} Q_A + \sqrt{c_B} Q_B)^2}. \quad (4.1)$$

Values less than 1.0 favor the screening approach while values greater than 1.0 favor the take-all approach.

We will illustrate efficiency using six scenarios regarding a survey of a hypothetical adult population. For all scenarios, the population size is taken from the March 2010 Current Population Survey (<http://www.census.gov/cps/data/>) and the population proportions by telephone status are obtained from the January – June 2010 National Health Interview Survey (Blumberg and Luke 2010). The values are  $N_A = 83,451,980$ ,  $N_a = 15,162,402$ ,  $N_{ab} = 68,289,578$ ,  $N_b = 31,265,108$ ,  $N_B = 99,554,686$ ,  $\alpha = 0.818$ , and  $\beta = 0.686$ . For all scenarios, the aim of the survey is taken to be the estimation of the total number of adults with a certain attribute.

The scenario specific assumptions are set forth in the following table:

The means correspond to the proportions of adults with the attribute. Scenario 1 describes a population in which the domain means are similar, with the mean of the dual-user domain being somewhat larger than the means of the CPO and LLO populations. Scenario 2 describes a population in which the mean of the LLO domain is somewhat larger than the means of the other telephone status domains. Scenario 3 reflects a population in which the means of all telephone status domains are equal. Scenario 4 reflects a population in which the mean of the LLO domain is much larger than the mean of the CPO domain. Scenarios 5 and 6 correspond to Scenarios 1 and 2, respectively, using means equal to one minus the corresponding means. The mean of the CPO domain declines from Scenario 1 to 6.

We selected the six scenarios to illustrate various circumstances in which the means of CPO, LLO, and dual-user domains differ. Differences can arise because younger adults, Hispanics, adults living only with unrelated adult roommates, adults renting their home, and adults living in poverty tend to be CPO (Blumberg and Luke 2013). To gain insight into the relative efficiencies of the take-all and screening designs, planners of future surveys may repeat our calculations for new scenarios specified by them and tailored to the particulars of their applications.

We will consider the six scenarios using three assumed cost structures. The cost structures are intended to illuminate various circumstances in which the per-unit cost of screening is high or low relative to the cost of the survey interview, with Cost Structures 1–3 reflecting increasing relative cost of screening. All cost components are expressed in interviewing hours:

Cost Structure 1:  $c'_B=0.05$ ,  $c''_B=2.05$ ,  $c_B=2.00$  and  $c_A=1.00$

Cost Structure 2:  $c'_B=0.20$ ,  $c''_B=2.20$ ,  $c_B=2.00$  and  $c_A=1.00$

Cost Structure 3:  $c'_B=0.50$ ,  $c''_B=2.50$ ,  $c_B=2.00$  and  $c_A=1.00$ .

All reflect circumstances in which the hours per case for a cell-phone interview is about 2 times larger than the hours per case for a landline interview.

Efficiencies corresponding to the various scenarios for the first cost structure are illustrated in Figure 4.1. We have prepared similar figures for the second and third cost structures, but to conserve space we do not present them here.

Given Cost Structure 1, the screening approach achieves the lower variance for the same fixed cost for all six scenarios. Given Cost Structure 3, in which the per-unit cost of screening is relatively much higher than in Cost Structure 1, the take-all approach achieves a smaller variance than the screening approach for half of the population scenarios. For Cost Structure 2, which entails an intermediate level of screening cost, the screening approach beats the take-all approach for all scenarios except for Scenario 1, in which the two approaches are nearly equally efficient.

The comparison between the take-all and screening protocols can be understood by examining the form of efficiency  $E$  in (4.1). The unit cost of screening is embedded only within the term  $\sqrt{c_B'''} R_B$  in the numerator of  $E$ . Thus, for a given scenario, the value of  $E$  must increase with increasing screening cost. For smaller screening costs,  $E$  may be less than 1.0 in which case the screening protocol will be preferred, while for larger screening costs,  $E$  may exceed 1.0 in which case the take-all protocol will be preferred.

It is also of interest to examine how the efficiency  $E$  varies with the domain means (i.e., the domain proportions), given a fixed cost structure. We see in (4.1) and in the definitions of the variance components that as long as the domain means –  $\bar{Y}_b$ ,  $\bar{Y}_{ab}$ , and  $\bar{Y}_a$  – vary reasonably together, as they do in our scenarios, the variation has relatively little or no impact on  $Q_A^2$ ,  $Q_B^2$ , and  $R_A^2$ , and  $E$  will tend to vary more directly with  $R_B^2$ , and in turn with the value of the ratio  $\bar{Y}_b^2/S_b^2$  in the CPO domain. The smaller the mean in the CPO domain, the smaller this ratio will be, and in turn the smaller  $E$  will be. Thus, in each of the structures, we see smaller values of  $E$  in Scenarios 5 and 6 than in Scenarios 1 and 2, and intermediate values of  $E$  in Scenarios 3 and 4.

For the take-all protocol, the optimum  $p$ 's are located at the points at which the efficiencies reach their maximum values. Table 4.2 reveals the optimum sample sizes and the optimum parameters  $p$  for each scenario and cost structure, assuming a fixed cost budget of 1,000 interviewing hours. For the screening protocol, we expect to complete  $(1 - \beta)n_B$  cell-phone interviews. For all population scenarios and cost structures studied here, the screening protocol obtains fewer completed cell-phone interviews than does the take-all protocol. The latter design uses resources for interviewing dual-user cases in both of the samples and requires more cell-phone interviews to provide adequate representation of CPO cases, while the former design can be more efficient about interviewing CPO cases at the price of using resources to conduct the requisite screening interviews. The optimum  $p$ 's fall approximately in the range from 0.4 to 0.6 and the variance under the take-all protocol is fairly flat within this range. We examine this issue further in Section 4.2.

In summary, one may conclude from these illustrations that the screening approach is often more efficient than the take-all approach. As the cost of the screener increases relative to the cost of the interview, the outcome can tip in favor of the take-all approach. The take-all approach will be preferred for surveys in which the cost of the screener is relatively very high; otherwise, the screening protocol will be preferred. The screening approach will tend to be relatively more efficient for small values of the CPO domain mean than for large values of this mean.

## 4.2 Choosing the mixing parameter $p$ for the take-all protocol

The optimum allocation is defined in terms of the mixing parameter, and thus it is important to consider the choice of this parameter. In the foregoing section, we saw that variance is likely not very sensitive to the choice of  $p$  within a reasonable neighborhood of optimum  $p$ . While the actual optimum  $p$  will never be known in practical applications, in this section, we



describe a practical method that statisticians may use to select a reasonable, near-optimum value of  $p$ .

The landline and cell-phone samples each supply an estimator of the total in the dual-user domain, and the mixing parameter  $p$  is used to combine the two estimators into one best estimator for this domain. When the estimator of the dual-user domain derived from the landline sample is the more precise,  $p$  should be relatively large, and conversely, when the estimator from the cell-phone sample is the more precise, then  $q = 1 - p$  should be relatively large. It makes good statistical sense to consider the value of  $p$  that is proportional to the expected sample size in the dual-user domain, i.e.,  $p_o = \alpha n_{A,opt} / (\alpha n_{A,opt} + \beta n_{B,opt})$ , where the optimum allocation is based on this choice of  $p$ . Thus,  $p_o$  is a root of the equation

$$\frac{c_A p^2}{c_B (1-p)^2} = \frac{(1-\alpha) S_a^2 + \alpha p^2 S_{ab}^2 + \alpha (1-\alpha) (\bar{Y}_a - p \bar{Y}_{ab})^2}{(1-\beta) S_b^2 + \beta (1-p)^2 S_{ab}^2 + \beta (1-\beta) \{\bar{Y}_b - (1-p) \bar{Y}_{ab}\}^2}, \quad (4.2)$$

and, in turn,  $n_{A,opt}$  and  $n_{B,opt}$  are defined in terms of  $p_o$ .

From (4.2) it is apparent that  $p_o$  is a function of the  $y$ -variable of interest. Use of this  $p_o$  in actual practice could imply a different sample size and set of survey weights for each variable of interest, which would be unworkable. To provide a practicable solution, one might consider use of the  $p_o$  that corresponds to the survey variable  $y \equiv 1$  (the population total corresponding to this variable is simply the total number of unique units on the two sampling frames). Given this approach  $p_o$  is a root of the equation

$$\frac{c_A p^2}{c_B (1-p)^2} = \frac{\alpha (1-\alpha) (1-p)^2}{\beta (1-\beta) p^2}. \quad (4.3)$$

For the cost structures considered in this section, the corresponding  $p_o$  is 0.52. In Figure 4.1, one can see that this value is very close to the exact optimum  $p$ 's under the various scenarios, with little loss in efficiency. Alternatively, one could evaluate (4.2) for a small set of the most important items in the survey; choose a good compromise value of  $p$ ; and then define the optimum allocation in terms of this one compromise value.

## 5 Example: *National Immunization Survey*

### 5.1 Introduction

CDC has sponsored the *National Immunization Survey* (NIS) since 1994 to monitor the vaccination status of young children age 19 – 35 months. The NIS uses two phases of data collection: a dual-frame RDD telephone survey of households with age-eligible children, followed by a mail survey of the vaccination providers of these children, which obtains vaccination histories for the children for each recommended vaccine. Each such child's provider-reported number of doses is compared to the recommended number of doses to

determine whether the child is up-to-date (UTD). Information about the NIS is available in Smith, Hoaglin, Battaglia, Khare and Barker (2005) and the 2011 Data User's Guide (CDC 2012).

We will discuss the NIS as it was conducted in 2011. The main interview consisted of six sections, beginning with Section S, which is a brief questionnaire module that determines whether the household has age-eligible children. The interview is then terminated for ineligible households. For eligible respondents with an available vaccination record (shotcard), Section A obtains the child(ren)'s household-reported vaccination history. For all other respondents, Section B obtains a more limited and less specific amount of information about the child(ren)'s vaccinations. Section C collects demographic characteristics of the child(ren), the mother, and the household. Section D collects the names and contact information for the child(ren)'s vaccination providers and requests parental consent to contact the providers, while Section E collects information regarding current health insurance coverage.

## 5.2 Optimum allocation for NIS

The NIS is designed to produce estimates at the national level and for 56 non-overlapping estimation areas, consisting of 46 whole states, 6 large urban areas, and 4 rest-of-state areas. Each of these areas is a sampling stratum in the NIS design. For each of these areas, NIS is designed to minimize the cost of the survey subject to a constraint on variance: the coefficient of variation (CV) of the estimator of the vaccination coverage rate (UTD children as a proportion of all eligible children) is to be 7.5 percent at the estimation-area level, when the true rate is 50 percent.

Given the take-all protocol, the six-part survey interview is administered to all respondents in both sample. Given the screening protocol, the survey interview is administered to all respondents in the landline sample, while in the cell-phone sample, the overall interview is now in two parts: (i) the brief screener to determine telephone status and (ii) the aforementioned six-part survey interview. Dual users are screened out of the cell-phone sample.

To illustrate the optimum allocation, we take the per-unit costs to be proportional to the following values:  $c'_B=0.06$ ,  $c''_B=2.03$ ,  $c_B=1.96$ , and  $c_A=1.00$ . Cell-phone interviews require roughly twice as many labor hours as landline interviews. We assume the following population proportions for age-eligible children by telephone status:  $W_A=0.59$ ,  $W_a=0.08$ ,  $W_{ab}=0.51$ ,  $W_b=0.41$ ,  $W_B=0.92$ ,  $\alpha=0.86$ , and  $\beta=0.55$ . We calculated these proportions using data from the January – June 2010 National Health Interview Survey.

To estimate a vaccination coverage rate given the take-all approach, we work with the variable

$$Y_i = \begin{cases} 1, & \text{if the } i^{\text{th}} \text{ case is an age-eligible child who is UTD} \\ 0, & \text{otherwise.} \end{cases}$$

Then, the estimated vaccination coverage rate is  $\bar{Y}/N_e$ , where  $N_e$  signifies the number of age-eligible children in the population (assumed known from vital statistics and related records). In accordance with the variance constraint, we take  $\bar{Y}_{ae} = \bar{Y}_{abe} = \bar{Y}_{be} = 0.5$ , where the subscript  $e$  signifies the mean of the age-eligible cases within the corresponding

telephone status domain. Then,  $\bar{Y}_d = \bar{Y}_{de}P_{de}$  and  $S_d^2 = \bar{Y}_d(1 - \bar{Y}_d)$ , where  $d = a, ab, b$  designates the three telephone status domains and  $P_{de} = N_{de}/N_d$  signifies the age-eligibility rate within domain  $d$ . Based on NIS experience, we take  $P_{ae} = 0.015$ ,  $P_{abe} = 0.03$ , and  $P_{be} = 0.05$ , reflecting an increasing eligibility rate across the telephone status domains; that is, young child-bearing families tend to have a cell phone and further tend to be CPO. By definition, the variance is the square of the coefficient of variation times the square of the population proportion. Thus, the variance constraint is  $\text{Var}\{\bar{Y}/N_e\} = 0.075^2 \times 0.5^2$ .

To estimate a vaccination coverage rate given the screening design, we work with the variable

$$Y_i = \begin{cases} 1, & \text{if } i \in s_A, \text{ and is an age- eligible child who is UTD} \\ 0, & \text{if } i \in s_A, \text{ and is not an age- eligible child or is not UTD} \\ 1, & \text{if } i \in s_B, \text{ and is CPO and is an age- eligible child who is UTD} \\ 0, & \text{if } i \in s_B, \text{ and is not CPO or is not an age- eligible child or is not UTD.} \end{cases}$$

Given these assumptions, the values of the efficiency ratio  $E$  lie below 1.0 for all values of  $p$  and from this we conclude that the screening design may be relatively less costly than the take-all design. The optimum value of  $p$  is about 0.39. However,  $E$  is quite flat in a neighborhood of the optimum and thus values of  $p$  in this neighborhood would produce similar total cost.

Given our assumptions, the optimum allocation for the take-all protocol at the optimum  $p$  is  $n_A = 3,069$  and  $n_B = 7,437$ , which equates to 86 NIS interviews on behalf of age-eligible children in the landline sample and 289 interviews on behalf of age-eligible children in the cell-phone sample. For the screening protocol, the optimum allocation is  $n_A = 5,858$  and  $n_B = 8,432$ , which we expect to yield 164 NIS interviews on behalf of age-eligible children in the landline sample and 188 NIS interviews of CPO households on behalf of their age-eligible children. These allocations apply to a single typical estimation area. Table 5.1 displays the expected sample sizes by telephone status domain given the optimum allocations. Given the screening protocol, the cell-phone sample yields an expected 4,674 dual users, which in turn reflect an expected 140 age-eligible children (who are not to be interviewed and thus are not included in the table).

We developed the optimum allocations revealed here under ideal conditions in which there is no nonresponse. To prepare a sample for actual use in the NIS (or any real survey), the allocation must be adjusted by the reciprocals of the expected survey cooperation rates and by the expected design effect due to weighting and clustering.

While the extant evidence shows that the screening protocol is slightly less costly than the take-all protocol, given that both achieve the same fixed variance constraint, the take-all

protocol actually provides the NIS an ongoing platform for testing and comparing both protocols. The authors continue to monitor the achieved sample composition and to conduct other specialized studies of response and nonresponse error.

## 6 Summary

We investigated two designs for a dual-frame telephone survey: a take-all protocol in which every respondent in the cell-phone sample is interviewed and a screening protocol in which respondents in the cell-phone sample are screened for phone status and only CPO respondents are interviewed. For each design, we derived the optimum allocation of the overall survey resources to the two sampling frames.

We studied the allocation problem given the two traditional meanings of the word “optimum”: (1) to minimize variance subject to a constraint on data collection cost, and (2) to minimize data collection cost subject to a constraint on variance. Given fixed variance, we find that the screening approach tends to achieve lower total cost than the take-all approach when the per-unit cost of screening is low relative the unit cost of the survey interview. The take-all approach can achieve the lower total cost when the per-unit cost of screening is relatively high. Similarly, given fixed total cost, the screening protocol tends to be the more efficient approach when the per-unit cost of screening is relatively low, and the take-all protocol can be the more efficient approach as the per-unit cost of screening rises. Both the landline and cell-phone samples have the capacity to produce estimators for the dual-user domain, while only the cell-phone sample can produce estimators for the CPO domain. Thus, when screening is relatively inexpensive on a per-unit basis, then it should be used to produce the largest possible sample from the CPO domain. But when screening is relatively expensive, then it is better to avoid the screening step and invest the survey resources in a larger interview sample. These results were obtained under an assumption of simple random sampling, and they may not carry over exactly to other sampling designs.

The take-all design results in two estimators for the dual-user domain, which are combined using factors of  $p$  and  $1 - p$  for the estimators from the landline and cell-phone samples, respectively. We studied the optimum choice of  $p$  and gave expressions for reasonable compromise values of  $p$ . When variance (or cost) is considered as a function of  $p$ , we found that it is fairly flat in a neighborhood of the optimum. The optimum allocation itself is a function of  $p$  and we found that the allocation is relatively insensitive to choices of  $p$  within a broad neighborhood of the optimum  $p$ .

We initiated this work before 2010 at a time when the CPO population in the U.S. was only a fifth to a quarter of the total population of households. At that time it made sense to contemplate a protocol in which the larger landline sample is interviewed in its entirety and the smaller cell-phone sample is screened for CPO status. At this writing, however, the CPO population comprises more than a third of the total population of households and it is still growing. It has become reasonable to consider a new screening protocol in which the landline sample is screened for telephone status and only LLO respondents are interviewed. The foregoing allocations and findings apply to this new protocol by symmetry.

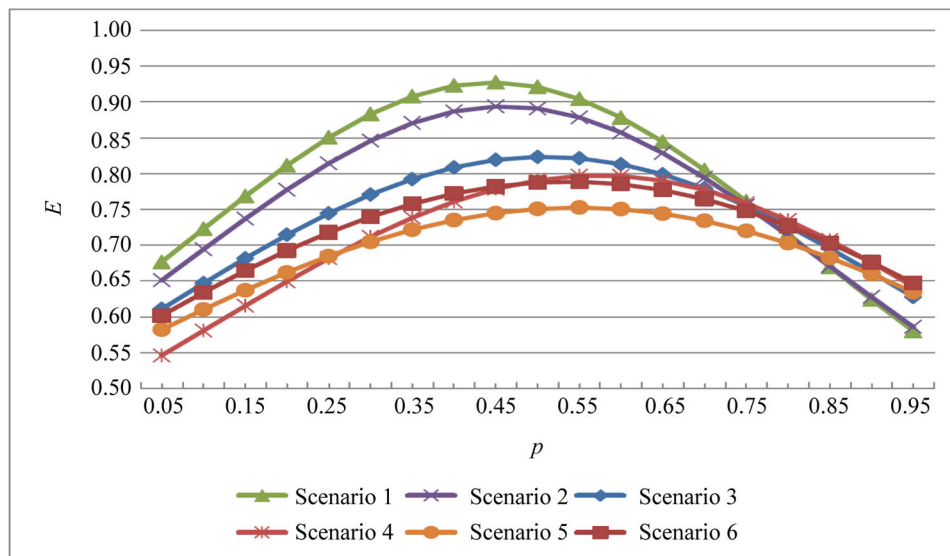
We illustrated the optimum allocations and the two interviewing protocols using the 2011 *National Immunization Survey*. The survey is designed to minimize cost under a fixed variance constraint. The NIS results are limited to the population of children age 19 – 35 months. Similar results may or may not obtain for a general population survey or for a survey with a different structure of per-unit costs.

## Acknowledgments

The authors' kindly acknowledge suggestions for improved readability offered by the Associate Editor and referees. Disclaimer: The findings and conclusions in this paper are those of the author(s), and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

## References

- Biemer, PP. Methodology for optimal dual frame sample design. Bureau of the Census SRD Research Report CENSUS/SRD/RR-84/07. 1984. available at [www.census.gov/srd/papers/pdf/rr84-07.pdf](http://www.census.gov/srd/papers/pdf/rr84-07.pdf)
- Blumberg, SJ., Luke, JV. Wireless substitution: Early release of estimates from the National Health Interview Survey, January–June 2010. National Center for Health Statistics; 2010 Dec. Available at <http://www.cdc.gov/nchs/nhis/releases.htm>
- Blumberg, SJ., Luke, JV. Wireless substitution: Early release of estimates from the National Health Interview Survey, July–December 2012. National Center for Health Statistics; 2013 Jun. Available at <http://www.cdc.gov/nchs/nhis/releases.htm>
- CDC. National Immunization Survey: A User's Guide for the 2011 Public Use Data File. 2012. Available at [http://www.cdc.gov/nchs/nis/data\\_files.htm](http://www.cdc.gov/nchs/nis/data_files.htm)
- Cochran, WG. Sampling Techniques. 3. New York: John Wiley & Sons, Inc; 1977.
- Fuller, WA., Burmeister, LF. Proceedings of the Social Statistics Section. American Statistical Association; 1972. Estimators for samples from two overlapping frames; p. 245-249.
- Hartley, HO. Proceedings of the Social Statistics Section. American Statistical Association; 1962. Multiple-frame surveys; p. 203-206.
- Hartley HO. Multiple frame methodology and selected applications. *Sankhyā*, Series C. 1974; 36:99–118.
- Lepkowski JM, Groves RM. A mean squared error model for multiple frame, mixed mode survey design. *Journal of the American Statistical Association*. 1986; 81:930–937.
- Lohr SL, Rao JNK. Inference from dual frame surveys. *Journal of the American Statistical Association*. 2000; 95:271–280.
- Lohr SL, Rao JNK. Estimation in multiple-frame surveys. *Journal of the American Statistical Association*. 2006; 101:1019–1030.
- Skinner CJ, Rao JNK. Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*. 1996; 91:349–356.
- Smith PJ, Hoaglin DC, Battaglia MP, Khare M, Barker LE. Statistical Methodology of the National Immunization Survey: 1994–2002. *National Center for Health Statistics, Vital and Health Statistics*. 2005; 2(138)
- Wolter KM, Smith P, Blumberg SJ. Statistical foundations of cell-phone surveys. *Survey Methodology*. 2010; 36(2):203–215.



**Figure 4.1.**  
Plot of efficiency  $E$  v. mixing parameter  $p$ , given cost structure 1.

Table 4.1

Definition of six scenarios for a hypothetical adult population

Scenarios	$\bar{Y}_A$	$\bar{Y}_a$	$\bar{Y}_{ab}$	$\bar{Y}_b$	$\bar{Y}_B$
1	0.791	0.750	0.800	0.750	0.784
2	0.759	0.800	0.750	0.750	0.750
3	0.500	0.500	0.500	0.500	0.500
4	0.518	0.600	0.500	0.400	0.469
5	0.209	0.250	0.200	0.250	0.216
6	0.241	0.200	0.250	0.250	0.250

**Table 4.2**

Sample sizes and optimum  $p^*$ 's for the take-all and screening designs

Cost Structure	Screening Design			Take-All Design		
	$n_A$	$n_B$	$(1 - \beta)n_B$	$p_{opt}$	$n_A$	$n_B$
Scenario 1						
1	494	747	234	0.45	337	331
2	469	641	201	0.45	337	331
3	431	505	159	0.45	337	331
Scenario 2						
1	506	728	229	0.45	339	330
2	481	626	197	0.45	339	330
3	443	494	155	0.45	339	330
Scenario 3						
1	583	615	193	0.50	344	328
2	559	533	167	0.50	344	328
3	520	425	134	0.50	344	328
Scenario 4						
1	605	582	183	0.55	377	312
2	581	506	159	0.55	377	312
3	543	405	127	0.55	377	312
Scenario 5						
1	606	581	182	0.55	358	321
2	582	505	159	0.55	358	321
3	544	404	127	0.55	358	321
Scenario 6						
1	618	563	177	0.55	354	323
2	594	490	154	0.55	354	323
3	557	393	123	0.55	354	323



**Table 5.1**

Expected sample sizes by telephone status domain given optimum allocations

Sample and Telephone Status Domains	Take-All Protocol		Screening Protocol	
	Expected Sample Size	Expected Age-Eligible Cases	Expected Sample Size	Expected Age-Eligible Cases
$s_A$	3,069	86	5,858	164
$s_B$	7,437	289	8,432	188
$s_A \cap U^a$	416	6	794	12
$s_A \cap U^{ab}$	2,653	80	5,064	152
$s_B \cap U^{ab}$	4,122	124	4,674	0
$s_B \cap U^b$	3,314	166	3,758	188